Parameter Identification in a Class of Linear Structural Equation Models

Jin Tian
Department of Computer Science
Iowa State University
Ames, IA 50011
jtian@cs.iastate.edu

Abstract

Linear causal models known as structural equation models (SEMs) are widely used for data analysis in the social sciences, economics, and artificial intelligence, in which random variables are assumed to be continuous and normally distributed. This paper deals with one fundamental problem in the applications of SEMs – parameter identification. The paper uses the graphical models approach and provides a procedure for solving the identification problem in a special class of SEMs.

1 Introduction

Linear causal models known as *structural equation models (SEMs)* are widely used for causal reasoning in the social sciences, economics, and artificial intelligence (AI) [Bollen, 1989; Pearl, 2000; Spirtes *et al.*, 2001]. In a SEM, the causal relationships among a set of variables are assumed to be linear and expressed by linear equations. As an example, consider the following model from [Pearl, 2000] that concerns with the relations between smoking (X) and lung cancer (Y), mediated by the amount of tar (Z) deposited in a person's lungs:

$$\begin{aligned} X &= \epsilon_1 \\ Z &= aX + \epsilon_2 \\ Y &= bZ + \epsilon_3 \\ Cov(\epsilon_1, \epsilon_2) &= Cov(\epsilon_2, \epsilon_3) = 0 \\ Cov(\epsilon_1, \epsilon_3) &\neq 0 \end{aligned}$$

The model makes the causal assumptions that the amount of tar Z deposited in the lungs depends on the level of smoking X (and external factors omitted from the model represented by ϵ_i assumed to have normal distribution) and that the production of lung cancer Y depends on the amount of tar in the lungs but smoking has no effect on lung cancer except as mediated through tar deposits. The external factors that have influence on smoking and cancer may be correlated (covariances $Cov(\epsilon_1, \epsilon_3) \neq 0$). The parameters a and b quantify the strength of linear cause-effect relationships.

SEMs are typically used for confirmatory data analysis in the social sciences and economics, consisting of four steps [Kenny *et al.*, 1998]: (1) hypothesizing a model, (2) identification analysis – to decide if there is a unique valuation for the parameters that make the model compatible with the observed data, (3) parameter estimation, and (4) evaluation of fit – to see how well the proposed model fits the data. In this paper, we will focus on the identification problem.

The identification problem has been under extensive study by econometricians and social scientists [Fisher, 1966; Bowden and Turkington, 1984; Bekker *et al.*, 1994; Rigdon, 1995]. In recent years the problem has been addressed using the graphical models techniques in the AI community [Pearl, 1998; Spirtes *et al.*, 1998; Tian, 2004]. A number of sufficient graphical criteria for identification have been developed, in [Brito and Pearl, 2002c; 2002b; 2002a; 2006] based on Wright's equations [Wright, 1934], and in [Tian, 2007a] using partial regression equations [Tian, 2005]. Most of these results are sufficient criteria which are applicable only when certain conditions are met.

Despite all this effort, the problem still remains open. In other words, we do not have a necessary and sufficient criterion for identification in arbitrary SEMs. One advancement in this direction is a necessary and sufficient procedure for identification in a special class of SEMs presented in [Tian, 2007b]. In this paper, we solve the identifiability problem in a class of SEMs strictly larger than those in [Tian, 2007b]. We present a procedure that will systematically determine whether each parameter in the model is identifiable or not and, if the answer is positive, the procedure will express the parameter in terms of observed covariances.

We begin with a formal introduction to SEMs and the identification problem, and introduce the partial regression equations method in [Tian, 2005] before presenting our results. For space reasons, the proofs are not included which can be found in the extended version of the paper.

2 Linear SEMs and Identification

A linear SEM over a set of random variables $V = \{V_1, \ldots, V_n\}$ is given by a set of structural equations of the form

$$V_j = \sum_i c_{ji} V_i + \epsilon_j, \quad j = 1, \ldots, n, \qquad (1)$$

where the summation is over the variables in V judged to be immediate causes of V_j. c_{ji}, called a *path coefficient*, quantifies the direct causal influence of V_i on V_j. ϵ_j's represent "error" terms and are assumed to have normal distri-

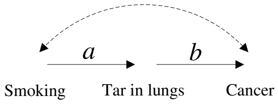


Figure 1: Path diagram illustrating the effect of smoking on lung cancer.

bution. In this paper we consider recursive models and assume that the summation in Eq. (1) is for $i < j$, that is, $c_{ji} = 0$ for $i \geq j$. The set of variables (and the corresponding structural equations) are considered to be ordered as $V_1 < V_2 < \ldots < V_n$. We denote the covariances between observed variables $\sigma_{ij} = Cov(V_i, V_j)$, and between error terms $\psi_{ij} = Cov(\epsilon_i, \epsilon_j)$. We denote the following matrices, $\Sigma = [\sigma_{ij}]$, $\Psi = [\psi_{ij}]$, and $C = [c_{ij}]$. Without loss of generality, the model is assumed to be standardized such that each variable V_j has zero mean and variance 1.

The structural assumptions encoded in a model are the zero path coefficients c_{ji} and zero error covariances ψ_{ij}. The model structure can be represented by a directed acyclic graph (DAG) G with (dashed) bidirected edges, called a *causal diagram* (or *path diagram*), as follows: the nodes of G are the variables $V_1, \ldots, V_n$; there is a directed edge from V_i to V_j in G if V_i appears in the structural equation for V_j, that is, $c_{ji} \neq 0$; there is a bidirected edge between V_i and V_j if the error terms ϵ_i and ϵ_j have non-zero correlation ($\psi_{ij} \neq 0$). For example, the smoking-and-lung-cancer SEM is represented by the causal diagram in Figure 1, in which each directed edge is annotated by the corresponding path coefficient.

The parameters of the model are the non-zero entries in the matrices C and Ψ. Fixing the model structure and given parameters C and Ψ, the covariance matrix Σ is given by (see, for example, [Bollen, 1989])

$$\Sigma = (I - C)^{-1}\Psi[(I - C)^{-1}]^t, \tag{2}$$

where "*t*" represents transpose. Conversely, in the identification problem, given the structure of a model, one attempts to solve for C in terms of the given observed covariance matrix Σ. If Eq. (2) gives a unique solution to some path coefficient c_{ji}, independent of the (unobserved) error correlations Ψ, the path coefficient c_{ji} is said to be *identified*; otherwise, c_{ji} is said to be *nonidentifiable*. In other words, the *identification problem* is that whether a path coefficient is determined uniquely from the covariance matrix Σ given the causal diagram. If every parameter of the model is identified, then *the model is identified*. Note that the identifiability conditions we seek involve the structure of the model alone, not particular numerical values of parameters, that is, we insist on having *identification almost everywhere*, allowing for pathological exceptions (see, for example, [Brito and Pearl, 2002a] for formal definition of identification almost everywhere).

3 Partial Regression Equations

In this paper, we will solve the identification problem using the partial regression equations method presented in [Tian, 2005] which will be introduced next.

For a set $S \subseteq V$, let $\beta_{ij.S}$ denote the *partial regression coefficient* which represents the coefficient of V_j in the linear regression of V_i on V_j and S. Note that partial regression coefficients can be expressed in terms of the covariance matrix Σ and that the order of the subscripts in $\beta_{ij.S}$ is essential. Let S_{jk} denote a set

$$S_{jk} = \{V_1, \ldots, V_{j-1}\} \setminus \{V_k\}. \tag{3}$$

[Tian, 2005] derived an expression for the partial regression coefficient $\beta_{jk.S_{jk}}$, for each pair of variables $V_k < V_j$, in terms of the model parameters (path coefficients and error covariances) given by

$$\beta_{jk.S_{jk}} = c_{jk} + \alpha_{jk} - \sum_{k+1 \leq l \leq j-1} \beta_{lk.S_{lk}} \alpha_{jl},$$
$$j = 2, \ldots, n, \; k = 1, \ldots, j-1, \tag{4}$$

where α_{jk}'s are defined recursively in terms of the error covariances as

$$\alpha_{jk} = \frac{Cov(\epsilon_j, \epsilon'_k)}{Cov(\epsilon'_k, \epsilon'_k)}, \tag{5}$$

where

$$\epsilon'_1 = \epsilon_1 \tag{6}$$

and

$$\epsilon'_j = \epsilon_j - \sum_{k=1}^{j-1} \alpha_{jk}\epsilon'_k, \; j = 2, \ldots, n. \tag{7}$$

For convenience, we will often use the shorthand notation $\beta_{jk.}$ to denote $\beta_{jk.S_{jk}}$.

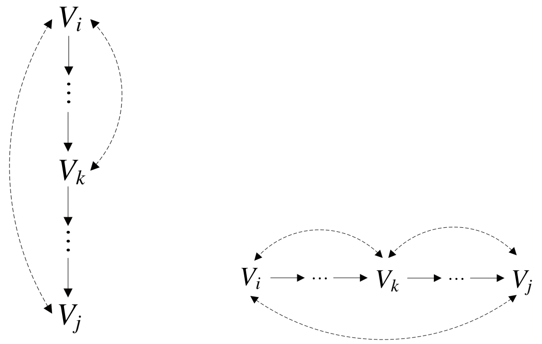
The set of equations given by (4) are called the *partial regression equations*. As an example, the partial regression equations for the model shown in Figure 1 are given by

$$\beta_{ZX} = a \tag{8}$$
$$\beta_{YZ.X} = b \tag{9}$$
$$\beta_{YX.Z} = \alpha_{YX}. \tag{10}$$

We immediately obtain that the path coefficients a and b are identified.

In general, given the model structure (represented by zero path coefficients and zero error correlations), some of the c_{jk} and α_{jk} will be set to zero in Eq. (4), and we can solve the identifiability problem by solving Eq. (4) for c_{jk} in terms of the partial regression coefficients. This provides a principled method for solving the identifiability problem. A path coefficient c_{ij} is identified if and only if the set of partial regression equations (4) give a unique solution to c_{ij}, independent of error correlations.

The partial regression equations are linear with respect to c_{jk}'s and α_{jk}'s, but may not be linear with respect to ψ_{ij}'s. The main difficulty in solving these equations lies in that α_{jk}'s are nonlinear functions of ψ_{ij}'s and may not be independent with each other. In this paper, we will study a class of SEMs in which we can treat α_{jk}'s as independent free parameters and thus for this class of SEMs the partial regression equations become linear equations.



(a) A ρ-structure. (b)A ρ-structure-free model.

Figure 2: Different structures of SEMs.

4 ρ-structure-free SEMs

[Tian, 2007b] studied so-called P-structure-free models which require that for any $i < k < j$ the two bidirected edges $V_j \leftrightarrow V_i$ and $V_k \leftrightarrow V_i$ can not both appear in the causal diagram (Fig. 2(a) and (b) both contain P-structures). In this paper, we relax this restriction and allow the simultaneous appearances of the two bidirected edges $V_j \leftrightarrow V_i$ and $V_k \leftrightarrow V_i$ as far as there also exists a bidirected edge between V_j and V_k.

Definition 1 (ρ-structure) *We will say that a SEM (or causal diagram) contains a ρ-*structure *if for some $i < k < j$, there is a bidirected edge between V_j and V_i, and a bidirected edge between V_i and V_k, but there is no bidirected edge between V_j and V_k (see Fig. 2a). Equivalently, in terms of model parameters, we say that a SEM contains a ρ-*structure *if for some $i < k < j$, $\psi_{ji} \neq 0$ and $\psi_{ki} \neq 0$ but $\psi_{jk} = 0$.*

We will say that a SEM (or causal diagram) is *ρ-structure-free* if it does not contain any ρ-structures. It is clear that P-structure-free SEMs are a strict subset of ρ-structure-free SEMs as any P-structure-free SEM is also ρ-structure-free but there exist models such as the one in Fig. 2(b) that is ρ-structure-free but is not P-structure-free.

In this paper we will study ρ-structure-free SEMs and show how to identify this class of models. First we show that in a ρ-structure-free SEM, α_{jk}'s can be treated as independent free parameters of the model.

Lemma 1 *In a ρ-structure-free SEM if $\psi_{jk} = 0$ then $\alpha_{jk} = 0$. Graphically speaking, if there is no bidirected edge between V_j and V_k, then $\alpha_{jk} = 0$.*

It is straightforward to show that *α_{jk}'s can be treated as independent parameters in place of ψ_{jk}'s.* Therefore, in ρ-structure-free SEMs, the set of partial regression equations (4) become linear in terms of the variables c_{jk} and α_{jk}. And the identification problem is reduced to that of solving (4) for c_{jk} in terms of the partial regression coefficients $\beta_{jk.}$, which leads to the following proposition.

Proposition 1 *In a ρ-structure-free SEM, a path coefficient c_{jk} is identified if and only if the set of linear equations (4) give a unique solution to c_{jk} that is independent of α_{jk}'s.*

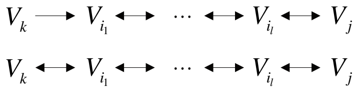


Figure 3: Two possible types of effective paths from V_k to V_j.

The difficulty of solving these linear equations lies in that the coefficients of these equations, the partial regression coefficients, are not independent parameters. The partial regression coefficients are related to each other in a complicated way, and it is difficult to decide the rank of the set of linear equations since it is not easy to determine whether certain expressions of partial regression coefficients will cancel out each other and become identically zero. To overcome this difficulty, next we show that the partial regression coefficients that appear in (4) can be expressed in terms of the free parameters c_{jk} and α_{jk}. First, we define some graphical notations.

A *path* between two nodes X and Y in a causal diagram consists of a sequence of consecutive edges of any type (directed or bidirected). A non-endpoint node Z on a path is called a *collider* if two arrowheads on the path meet at Z, i.e. $\rightarrow Z \leftarrow$, $\leftrightarrow Z \leftrightarrow$, $\leftrightarrow Z \leftarrow$, $\rightarrow Z \leftrightarrow$; all other non-endpoint nodes on a path are *non-colliders*, i.e. $\leftarrow Z \rightarrow$, $\leftarrow Z \leftarrow$, $\rightarrow Z \rightarrow$, $\leftrightarrow Z \rightarrow$, $\leftarrow Z \leftrightarrow$.

Definition 2 (Effective Path) *Let $k < j$. A path $(V_k, V_{i_1}, \ldots, V_{i_l}, V_j)$ from V_k to V_j is said to be an* effective path *if every intermediate node on the path is a collider, and $k < i_1 < \ldots < i_l < j$ (see Figure 3).*

We assume that the edges in the causal diagram are associated with the model parameters as follows:

- each directed edge $V_j \leftarrow V_k$ is associated with the path coefficient c_{jk}.
- each bidirected edge $V_j \leftrightarrow V_k$ where $k < j$ is associated with the parameter α_{jk}.

For a path p, let $T(p)$ *represent the product of the parameters along path p.* For example, let p be the path $V_1 \rightarrow V_2 \rightarrow V_3 \leftrightarrow V_5$ in Figure 4(b). Then $T(p) = c_{21}c_{32}\alpha_{53}$.

Lemma 2 *In a ρ-structure-free SEM,*

$$\beta_{jk.} = \sum_{p:\text{effective paths}} (-1)^{|p|-1} T(p), \tag{11}$$

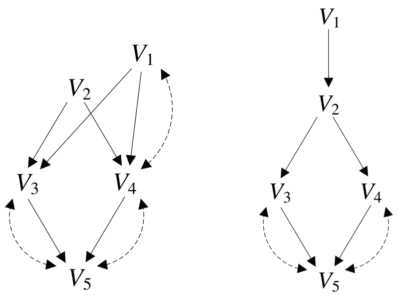
in which the summation is over all the effective paths from V_k to V_j and $|p|$ represents the number of edges on p.

As a corollary of Lemma 2 we have that $\beta_{jk.} = 0$ if there is no effective paths from V_k to V_j.

Next, we show how to solve the set of partial regression equations (4) in a ρ-structure-free SEM.

5 Identifying ρ-structure-free SEMs

According to Eq. (4), to decide the identifiability of a path coefficient c_{jk}, we need to solve the $j-1$ equations in (4) for $k = 1, \ldots, j-1$ simultaneously with c_{jl} and α_{jl} for $l < j$ as variables. And c_{jk} is identified if and only if the set of $j-1$



(a) (b)

Figure 4: Example SEMs.

equations give a unique solution to c_{jk} in terms of $\beta_{il.}$'s. For convenience, we will name the equation for $\beta_{jk.}$ after V_k (for a fixed j) as follows:

$$(V_k): \beta_{jk.} = c_{jk} + \alpha_{jk} - \sum_{k+1\le l\le j-1} \beta_{lk.}\alpha_{jl}. \tag{12}$$

Assuming that there is a directed edge $V_k \rightarrow V_j$ in the causal diagram, the path coefficient c_{jk} only appears once in this $j-1$ equations, that is, in the equation (V_k). Let PA_j be the set of parents of V_j (that is, the set of variables V_k such that $c_{jk} \neq 0$). Let $E(S)$ *denote the set of equations* (V_k) *such that* $V_k \in S$. Each equation (V_k) in $E(PA_j)$ can be solved for the path coefficient c_{jk} by simply rewriting the equation to obtain

$$(V_k): c_{jk} = \beta_{jk.} - \alpha_{jk} + \sum_{k+1\le l\le j-1} \beta_{lk.}\alpha_{jl}, \quad V_k \in PA_j. \tag{13}$$

Therefore c_{jk} is identifiable if none of the α_{ji} appears in this equation or all the α_{ji} appearing in the equation are identifiable. The problem of identifying c_{jk} is reduced to the problem of identifying α_{ji}'s.

Let $V_j^{<} = \{V_1, \ldots, V_{j-1}\}$ denote the set of variables ordered ahead of V_j. Let $\overline{PA}_j = V_j^{<} \setminus PA_j$. To identify α_{ji}'s we need to solve the set of equations in $E(\overline{PA}_j)$ with α_{ji}'s as unknowns, rewritten in the following:

$$(V_k): \beta_{jk.} = \alpha_{jk} - \sum_{k+1\le l\le j-1} \beta_{lk.}\alpha_{jl}, \quad V_k \in \overline{PA}_j. \tag{14}$$

Let SP_j (the set of spouses of V_j) be the set of variables V_k that connects with V_j by a bidirected edge $V_k \leftrightarrow V_j$ (that is, $\psi_{jk} \neq 0$). Then the number of unknowns is given by the number of variables in SP_j, denoted by $|SP_j|$. In general we may have more equations than unknowns ($|\overline{PA}_j| \geq |SP_j|$), or more unknowns than equations ($|\overline{PA}_j| \leq |SP_j|$). And these equations may not be linearly independent with each other.

For example, assume that we are interested in identifying the path coefficients c_{53} and c_{54} in the model in Figure 4(a). The set of equations $E(PA_j)$ in (13) become

$$(V_3): c_{53} = \beta_{53.124} - \alpha_{53} \tag{15}$$

$$(V_4): c_{54} = \beta_{54.123} - \alpha_{54}. \tag{16}$$

And the set of equations $E(\overline{PA}_j)$ become

$$(V_1): \beta_{51.234} = -\beta_{31.2}\alpha_{53} - \beta_{41.23}\alpha_{54} \tag{17}$$

$$(V_2): \beta_{52.134} = -\beta_{32.1}\alpha_{53} - \beta_{42.13}\alpha_{54}. \tag{18}$$

(V_1) and (V_2) may be solved simultaneously to identify α_{53} and α_{54} (almost everywhere), and therefore c_{53} and c_{54} are identified. On the other hand, in the model in Figure 4(b), the set of equations $E(\overline{PA}_j)$ become

$$(V_1): \beta_{51.234} = 0 \tag{19}$$

$$(V_2): \beta_{52.134} = -\beta_{32.1}\alpha_{53} - \beta_{42.13}\alpha_{54}. \tag{20}$$

We obtain that α_{53} and α_{54} are not identified.

In general, to solve the set of linear equations $E(\overline{PA}_j)$ in (14), we look for linearly independent equations. Next we show that this can be achieved by solving a maximum flow problem. We acknowledge that the idea of using the maximum flow technique was proposed by [Brito and Pearl, 2002b] and also used in [Tian, 2007a]. Still, constructing a relevant flow network poses a nontrivial problem.

5.1 Flow network

A flow network $F = (V, E)$ is a directed graph in which each edge $(u, v) \in E$ has a nonnegative capacity $c(u, v) \geq 0$ (see, for example, [Cormen *et al.*, 1990]). We distinguish two vertices in a flow network: a source s and a sink t. A flow in F is a real-valued function $f : V \times V \rightarrow R$ that satisfies the capacity constraints $f(u, v) \leq c(u, v)$ and the flow conservation property (the amount of flow entering any vertex must be the same as the amount of flow leaving the vertex) among others. The value of a flow f is defined as $|f| = \sum_{v \in V} f(s, v)$, that is, the total net flow out of the source. In the maximum flow problem, we are given a flow network F, with source s and sink t, and we wish to find a flow of maximum value from s to t.

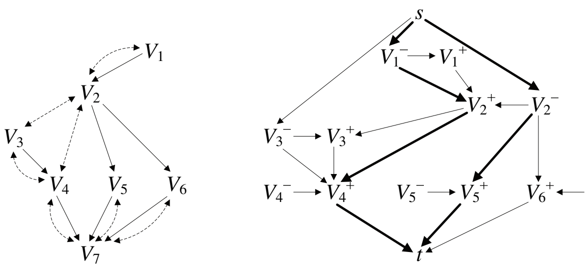
To facilitate identifying a set of linearly independent equations in $E(\overline{PA}_j)$, we construct a flow network F_j as follows. The nodes of F_j consists of:

- for every node $V_i < V_j$, add two nodes V_i^- and V_i^+ into F_j.
- a source node s.
- a sink node t.

The edges of F_j are:

- for every node $V_i < V_j$, add edge $V_i^- \rightarrow V_i^+$.
- for every edge $V_i \rightarrow V_l$, add edge $V_i^- \rightarrow V_l^+$.
- for every edge $V_i \leftrightarrow V_l,\ i < l$, add edge $V_i^+ \rightarrow V_l^+$.
- for every node $V_i \in SP_j$ (those with $\alpha_{V_j V_i} \neq 0$), add edge $V_i^+ \rightarrow t$.
- for every node $V_i \in \overline{PA}_j$ (those with $c_{ji} = 0$), add $s \rightarrow V_i^-$.

We assign a capacity 1 to every edge in F_j. We also assign a node capacity of 1 to every node (except s and t) in F_j (this can be achieved by splitting every node into two and connecting them by an edge of capacity 1 [Even, 1979]). As an example, for the model shown in Figure 5(a), the flow network relative to V_7 is given in Figure 5(b).



(a) (b)

Figure 5: A SEM and corresponding flow network.

Assume that we have solved the maximum flow problem on the flow network F_j (using, for example, Ford-Fulkerson algorithm). Since every edge has a capacity 1 and every node has a capacity 1, the computed flow f represents a set of disjoint directed paths from s to t. Let the set of directed paths be

$$q_i = s \rightarrow Z_i^- \rightarrow \ldots \rightarrow X_i^+ \rightarrow t,\ i = 1, \ldots, k,$$

where $k = |f|$. From the network construction, we have that $Z_i \in \overline{PA}_j$ and $X_i \in SP_j$. We will call the set of variables $AC = \{Z_1, \ldots, Z_k\} \subseteq \overline{PA}_j$ a *contributory set* relative to $DC = \{X_1, \ldots, X_k\} \subseteq SP_j$ and V_j. For example, the flow network in Figure 5(b) shows a maximum flow solution, which identifies the set $\{V_1, V_2\}$ as a contributory set relative to $\{V_4, V_5\}$ and V_7.

A contributory set plays a key role in solving the set of equations $E(\overline{PA}_j)$. Using Lemma 2 and the properties of the maximum flow we obtain the following theorems.

Theorem 1 *In a ρ-structure-free SEM, let $AC \subseteq \overline{PA}_j$ be a contributory set relative to $DC \subseteq SP_j$ and V_j, then the set of linear equations $E(AC)$ are linearly independent with respect to the set of unknowns $\alpha(DC) = \{\alpha_{V_j X_i} | X_i \in DC\}$, that is, $E(AC)$ can be solved with respect to unknowns in $\alpha(DC)$ (almost everywhere).*

The rest of the equations in $E(\overline{PA}_j)$ will be linearly dependent on the set of equations in $E(AC)$.

Theorem 2 *The number of linearly independent equations in $E(\overline{PA}_j)$ (that is, the rank of the coefficient matrix of the equations) is equal to the maximum flow value $|f|$.*

5.2 Solving the problem

From Theorems 1 and 2, after we have identified a contributory set AC relative to V_j we can determine the identifiability of the path coefficient c_{jk} by solving the set of equations $E(PA_j)$ and $E(AC)$. The following is a direct corollary of Theorems 1 and 2.

Theorem 3 *In a ρ-structure-free SEM, a path coefficient c_{jk} is identified if and only if the set of linearly independent equations $E(PA_j)$ and $E(AC)$ give a unique solution to c_{jk} that is independent of α_{jk}'s.*

In general, we can first solve the set of linear equations $E(AC)$ for unknown variables in $\alpha(DC)$ in terms of $\beta_{jk.}$'s

For $j = 1, \ldots, n$,

1. *Express c_{jk}'s in terms of parameter α_{jk}'s using the equations in $E(PA_j)$.*
2. *Construct the flow network for V_j.*
3. *Solve the maximum flow problem to obtain a contributory set AC_j relative to DC_j and V_j.*
4. *Solve $E(AC_j)$ with respect to the unknown variables $\{\alpha_{jk} | V_k \in DC_j\}$.*
5. *Substitute the values of solved variables into the equations $E(PA_j)$ to determine the identifiability of the path coefficients c_{jk}'s.*

Figure 6: A procedure for systematically identifying the path coefficients in a ρ-structure-free SEM.

and possibly some $\alpha_{V_j X_i}$'s for $X_i \notin DC$. Then we substitute the values of solved variables into the equations $E(PA_j)$ given in (13) to determine the identifiability of the path coefficients c_{jk}'s.

In summary, a procedure for systematically identifying the path coefficients in a ρ-structure-free SEM is given in Figure 6. For $j = 1, \ldots, n$, at each step, we attempt to identify parameters associated with the variable V_j. The procedure will tell which c_{jk}'s are identifiable, and which are not.

Notice that the equations in $E(\overline{PA}_j \setminus AC)$ are linearly dependent on the equations in $E(AC)$ and therefore are not useful for determining the identifiability of parameters. If we substitute the values of solved variables in $\alpha(DC)$ into the equations in $E(\overline{PA}_j \setminus AC)$, we will get a set of equations involving $\beta_{il.}$'s. These equations represent the set of constraints on the covariance matrix implied by the model structure. They can be used for testing a hypothesized model against the observed data.

5.3 An example

We illustrate the identification procedure by an example. Consider the model in Figure 5(a). Assume that we want to identify the path coefficients associated with V_7 (c_{74}, c_{75}, and c_{76}). First we express the path coefficients in terms of α_{7i}'s as follows

$$(V_6) : c_{76} = \beta_{76.12345} - \alpha_{76}. \tag{21}$$
$$(V_5) : c_{75} = \beta_{75.12346} - \alpha_{75}. \tag{22}$$
$$(V_4) : c_{74} = \beta_{74.12356} - \alpha_{74}. \tag{23}$$

Then we construct the flow network shown in Figure 5(b) and solve the maximum flow problem. Assume that the solution returns $\{V_1, V_2\}$ as a contributory set relative to $\{V_4, V_5\}$. Then we solve the equations (V_1) and (V_2) given in the following

$$(V_1) : \beta_{71.23456} = -\beta_{41.23}\alpha_{74} \tag{24}$$
$$(V_2) : \beta_{72.1345} = -\beta_{42.13}\alpha_{74} - \beta_{52.134}\alpha_{75} - \beta_{62.1345}\alpha_{76} \tag{25}$$

with α_{74} and α_{75} as unknown variables. We obtain

$$\alpha_{74} = -\beta_{71.23456}/\beta_{41.23} \tag{26}$$

$$\alpha_{75} = -\beta_{72.1345}/\beta_{52.134} + \beta_{42.13}\beta_{71.23456}/(\beta_{41.23}\beta_{52.134}) - \beta_{62.1345}\alpha_{76}/\beta_{52.134} \tag{27}$$

We conclude that α_{74} is identified, and that α_{75} and α_{76} are nonidentifiable. Finally, we substitute expressions for α_{74} and α_{75} into Eqs. (21)-(23), and we conclude that c_{74} is identified, and c_{75} and c_{76} are both nonidentifiable.

We notice that we have not used equation (V_3) given below

$$(V_3) : \beta_{73.12456} = -\beta_{43.12}\alpha_{74}, \tag{28}$$

which is indeed linearly dependent on the equations (V_1) and (V_2). In fact if we substitute into (V_3) the solved value of α_{74} given in Eq. (26) we obtain

$$\beta_{41.23}\beta_{73.12456} = \beta_{43.12}\beta_{71.23456}, \tag{29}$$

which represents a constraint on the covariance matrix imposed by the model structure.

6 Conclusion and Discussion

The identification problem has been a long standing problem in the applications of linear SEMs. Given a SEM, we would like to know which parameters in the model are uniquely determined by the observed covariances and which parameters are not, and we would like to know what constraints are implied by the model structure on the covariance matrix. In this paper, we provide a procedure for answering these questions in the class of ρ-structure-free SEMs.

In related work using graphical models methods, sufficient criteria for *model identification* have been developed in [Brito and Pearl, 2002c; 2002b; 2006], which established sufficient conditions for *all* the parameters in the model to be identified but can not be used to identify individual parameters if there exist nonidentifiable parameters in the model. A number of sufficient criteria for identifying individual parameters have been developed in [Pearl, 2000; Brito and Pearl, 2002a; Tian, 2007a]. Given a model, these methods may identify certain parameters but make no claims about other parameters.

The closest related work is a necessary and sufficient procedure for identifying P-structure-free SEMs [Tian, 2007b]. The ρ-structure-free SEMs we have solved in this paper contain P-structure-free models as a strict subset. The ultimate goal of this line of research is to provide a necessary and sufficient algorithm for identifying any possible models that may be hypothesized by researchers using SEMs. We believe this work is an important advance in this direction as there exist a large number of possible models containing the structure in Fig. 2(b) that are ρ-structure-free but not P-structure-free.

Acknowledgments

This research was partly supported by NSF grant IIS-0347846.

References

[Bekker *et al.*, 1994] P.A. Bekker, A. Merckens, and T.J. Wansbeek. *Identification, equivalent models, and computer algebra*. Academic, 1994.

[Bollen, 1989] K.A. Bollen. *Structural Equations with Latent Variables*. John Wiley, New York, 1989.

[Bowden and Turkington, 1984] R.J. Bowden and D.A. Turkington. *Instrumental Variables*. Cambridge University Press, Cambridge, England, 1984.

[Brito and Pearl, 2002a] C. Brito and J. Pearl. Generalized instrumental variables. In *Proceedings of the UAI*, 2002.

[Brito and Pearl, 2002b] C. Brito and J. Pearl. A graphical criterion for the identification of causal effects in linear models. In *Proceedings of the AAAI*, 2002.

[Brito and Pearl, 2002c] C. Brito and J. Pearl. A new identification condition for recursive models with correlated errors. *Structural Equation Modelling*, 9(4):459–474, 2002.

[Brito and Pearl, 2006] C. Brito and J. Pearl. Graphical condition for identification in recursive sem. In *Proceedings of the UAI*, 2006.

[Cormen *et al.*, 1990] Thomas H. Cormen, Charle E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. The MIT Press, 1990.

[Even, 1979] S. Even. *Graph Algorithms*. Computer Science Press, Rockville, Md, 1979.

[Fisher, 1966] F.M. Fisher. *The Identification Problem in Econometrics*. McGraw-Hill, 1966.

[Kenny *et al.*, 1998] D.A. Kenny, D.A. Kashy, and N. Bolger. Data analysis in social psychology. In D. Gilbert, S. Fiske, and G. Lindzey, editors, *The Handbook of Social Psychology*, pages 233–265. McGraw-Hill, 1998.

[Pearl, 1998] J. Pearl. Graphs, causality, and structural equation models. *Socioligical Methods and Research*, 27:226–284, 1998.

[Pearl, 2000] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, NY, 2000.

[Rigdon, 1995] E.E. Rigdon. A necessary and suficient identification rule for structural models estimated in practice. *Multivariate Behavioral Research*, 30:359–383, 1995.

[Spirtes *et al.*, 1998] P. Spirtes, T. Richardson, C. Meek, R. Scheines, and C. Glymour. Using path diagrams as a structural equation modeling tool. *Socioligical Methods and Research*, 27:182–225, 1998.

[Spirtes *et al.*, 2001] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search (2nd Edition)*. MIT Press, Cambridge, MA, 2001.

[Tian, 2004] J. Tian. Identifying linear causal effects. In *Proceedings of the AAAI*, 2004.

[Tian, 2005] J. Tian. Identifying direct causal effects in linear models. In *Proceedings of the AAAI*, 2005.

[Tian, 2007a] J. Tian. A criterion for parameter identification in structural equation models. *Proceedings of UAI*, 2007.

[Tian, 2007b] J. Tian. On the identification of a class of linear models. In *Proceedings of the AAAI*, 2007.

[Wright, 1934] S. Wright. The method of path coefficients. *Ann. Math. Statist.*, 5:161–215, 1934.